Bucketing Coding and Information Theory for the Statistical High Dimensional Nearest Neighbor Problem

Moshe Dubiner

ACKNOWLEDGMENT

This paper is dedicated to my wife Edith whose support made it possible, and to Benjamin Weiss who taught me what mathematical information means. I also thank Uri Zwick for suggesting clarifications.

M. Dubiner is with Google, e-mail: moshe@google.com

Manuscript submitted to IEEE Transactions on Information Theory on March 3 ,2007; revised August 27, 2007.

Bucketing Coding and Information Theory for the Statistical High Dimensional Nearest Neighbor Problem

Abstract

Consider the problem of finding high dimensional approximate nearest neighbors, where the data is generated by some known probabilistic model. We will investigate a large natural class of algorithms which we call bucketing codes. We will define bucketing information, prove that it bounds the performance of all bucketing codes, and that the bucketing information bound can be asymptotically attained by randomly constructed bucketing codes.

For example suppose we have n Bernoulli(1/2) very long (length $d \to \infty$) sequences of bits. Let $n - 2m$ sequences be completely independent, while the remaining $2m$ sequences are composed of m independent pairs. The interdependence within each pair is that their bits agree with probability $1/2 < p \leq 1$. It is well known how to find most pairs with high probability by performing order of $n^{\log_2 2/p}$ comparisons. We will see that order of $n^{1/p+\epsilon}$ comparisons suffice, for any $\epsilon > 0$. Moreover if one sequence out of each pair belongs to a a known set of $n^{(2p-1)^2-\epsilon}$ sequences, than pairing can be done using order n comparisons!

I. INTRODUCTION

Suppose we have two bags of points, X_0 and X_1, randomly distributed in a high-dimensional space. The points are independent of each other, with one exception: there is one unknown point x_0 in bag X_0 that is significantly closer to an unknown point x_1 in bag X_1 than would be accounted for by chance. We want an efficient algorithm for quickly finding these two 'paired' points. More generally, one could have m special pairs (up to having all points paired). An algorithm that finds a single pair with probability S will find an expected number of mS pairs, so keeping m as a parameter is unnecessary.

We worked on finding texts that are translations of each other, which is a two bags problem (the bags are languages). In most cases there is only one bag $X_0 = X_1 = X$, $n_0 = n_1 = n$.

The two bags model is slightly more complicated, but leads to clearer thinking. It is a bit reminiscent of fast matrix multiplication: even when one is interested only in square matrices, it pays to consider rectangular matrices too.

Let us start with the well known simple uniform marginally Bernoulli(1/2) example. Suppose $X_0, X_1 \subset \{0,1\}^d$ of sizes n_0, n_1 respectively are randomly chosen as independent Bernoulli(1/2) variables, with one exception. Choose uniformly randomly one point $x_0 \in X_0$, xor it with a random Bernoulli(p) vector and overwrite one uniformly chosen random point $x_1 \in X_1$. A symmetric description is to say that x_0, x_1 i'th bits have the joint probability matrix

$$P = \begin{pmatrix} p/2 & (1-p)/2 \\ (1-p)/2 & p/2 \end{pmatrix} \tag{1}$$

for some known $1/2 < p \leq 1$. In practice p will have to be estimated.

Let

$$\ln N = \ln n_0 + \ln n_1 - I(P)d \tag{2}$$

where

$$I(P) = I(p) = p\ln(2p) + (1-p)\ln(2(1-p)) \tag{3}$$

is the mutual information between the special pair's single coordinate values. Information theory tells us that we can not hope to pin the special pair down into less than N possibilities, but can come close to it in some asymptotic sense. Assume that N is small. How can we find the closest pair? The trivial way to do it is to compare all the $n_0 n_1$ pairs. A better way has been known for a long time. The earliest references I am aware of are Karp,Waarts and Zweig [7], Broder [3], Indyk and Motwani [6]. They do not limit themselves to this simplistic problem, but their approach clearly handles it. Without restricting generality let $n_0 \leq n_1$. Randomly choose

$$k \approx \log_2 n_0 \tag{4}$$

out of the d coordinates, and compare the point pairs which agree on these coordinates (in other

words, fall into the same bucket). The expected number of comparisons is

$$n_0 n_1 2^{-k} \approx n_1 \tag{5}$$

while the probability of success of one comparison is p^k. In case of failure try again, with other random k coordinates. At first glance it might seem that the expected number of tries until success is p^{-k}, but that is not true because the attempts are interdependent. An extreme example is $d = k$, where the attempts are identical. In the unlimited data case $d \to \infty$ the expected number of tries is indeed p^{-k}, so the expected number of comparisons is

$$W \approx p^{-k} n_1 \approx n_0^{\log_2 1/p} n_1 \tag{6}$$

Is this optimal? Alon [1] has suggested the possibility of improvement by using Hamming's perfect code.

We have found that in the $n_0 = n_1 = n$ case, $W \approx n^{\log_2 2/p}$ can be reduced to

$$W \approx n^{1/p+\epsilon} \tag{7}$$

for any $1/2 < p < 1$, $\epsilon > 0$. This particular algorithm is described in the next section. Amazingly it is possible to characterize the asymptotically best exponent not only for this problem, but for a much larger class. We allow non binary discrete data, a limited amount of data ($d < \infty$) and a general probability distribution of each coordinate.

We will prove theorem 10.1, a lower bound on the work performed by any bucketing algorithm. It employs a newly defined **bucketing information** function $I(P, \lambda_0, \lambda_1, \mu)$, which generalizes Shannon's mutual information function $I(P) = I(P, 1, 1, \infty)$. Comparing (2) with theorem 10.1 shows that the mutual information's function generalizes as well. Bucketing algorithms approaching the information bound are constructed by random coding. The analogy with Shannon's coding and information theory is very strong, suggesting that maybe we are redoing it in disguise. If it is a disguise, it is quite effective. Coding with distortion theory seems also related. There is related work [9], which tackles a particular class of practical bucketing algorithms (lexicographic forest algorithms). Their performance turns out to be bounded by a **bucketing forest information**

function, and that bound is asymptotically attained by a specific practical algorithm.

II. An Asymptotically Better Algorithm

The following algorithm does not generalize well, but makes sense for the uniform marginally Bernoulli(1/2) problem (1) with $1/2 < p < 1$. Let $0 < d_0 \le d$ be some natural numbers. We construct a d dimensional bucket in the following way. Choose a random point $b \in \{0,1\}^d$. The bucket contains all points $x \in \{0,1\}^d$ such for exactly $d_0 - 1$ or d_0 coordinates i $x_i = b_i$. (It is even better to allow $d_0 - 1, \ldots, d$, but the analysis gets a little messy.) The algorithm uses T such buckets, independently chosen. The probability of a point x falling into a bucket is

$$p_{A*} = \binom{d}{d_0 - 1} 2^{-d} + \binom{d}{d_0} 2^{-d} \tag{8}$$

Let the number of points be

$$n_0 = n_1 = n = \lfloor 1/p_{A*} \rfloor \tag{9}$$

This way the expected number of comparisons (point pairs in the same bucket) is

$$T(np_{A*})^2 \le T \tag{10}$$

The probability that both special pair points fall at least once into the same bucket is

$$S = \sum_{m=0}^{d} \binom{d}{m} p^{d-m}(1-p)^m \left[1 - (1 - S_m)^T\right] \tag{11}$$

$$S_m = 2^{-d} \binom{m}{\lfloor m/2 \rfloor} \left[\binom{d-m}{d_0 - \lceil m/2 \rceil} + \binom{d-m}{d_0 - \lceil (m+1)/2 \rceil}\right] \tag{12}$$

The explanation follows. In these formulas m is the number of coordinates i at which the special pair values disagree: $x_{0,i} \ne x_{1,i}$. Consider the special pair fixed. There are 2^d possible baskets, independently chosen. Consider one basket. For $j,k = 0,1$ denote by m_{jk} the number of coordinates i such that $x_{0,i} \oplus b_i = j$ and $x_{0,i} \oplus x_{1,i} = k$ where $\oplus$ is the xor operation. We know that $m_{01} + m_{11} = m$ and $m_{00} + m_{10} = d - m$. Both x_0, x_1 fall into the basket iff

$m_{00}+m_{01}=d_0-1,d_0$ and $m_{00}+m_{11}=d_0-1,d_0$. There are two possibilities

$$\begin{pmatrix} m_{00} & m_{01} \\ m_{10} & m_{11} \end{pmatrix} = \begin{pmatrix} d_0-\lceil m/2\rceil & \lfloor m/2\rfloor \\ d-d_0-\lfloor m/2\rfloor & \lceil m/2\rceil \end{pmatrix} \tag{13}$$

$$\begin{pmatrix} m_{00} & m_{01} \\ m_{10} & m_{11} \end{pmatrix} = \begin{pmatrix} d_0-\lceil (m+1)/2\rceil & \lceil m/2\rceil \\ d-d_0-\lfloor (m-1)/2\rfloor & \lfloor m/2\rfloor \end{pmatrix} \tag{14}$$

each providing

$$\begin{pmatrix} m_{00}+m_{10} \\ m_{00} \end{pmatrix} \begin{pmatrix} m_{01}+m_{11} \\ m_{01} \end{pmatrix} \tag{15}$$

buckets.

Clearly m obeys a Bernoulli$(1-p)$ distribution, so by Chebyshev's inequality

$$S \geq \min_{|m-(1-p)d|<\sqrt{p(1-p)d/\epsilon}} \left(1-e^{-TS_m}-\epsilon\right) \tag{16}$$

for any $0<\epsilon<1$. Hence taking

$$T=\lceil -\ln\epsilon / \min_{|m-(1-p)d|<\sqrt{p(1-p)d/\epsilon}} S_m \rceil \tag{17}$$

guaranties a success probability $S \geq 1-2\epsilon$. What is the relationship between n and T? Let

$$d_0 \sim (1+\rho)d/2, \quad d\to\infty \tag{18}$$

By Stirling's approximation

$$\lim \frac{\ln n}{d} = I\left(\frac{1+\rho}{2}\right) \tag{19}$$

$$\lim \frac{\ln T}{d} = pI\left(\frac{1+\rho/p}{2}\right) \tag{20}$$

Letting $\rho\to 0$ results in exponent

$$\lim \frac{\ln T}{\ln n} = \frac{1}{p} \tag{21}$$

We are not yet finished with this algorithm, because the number of comparisons is not the only component of work. One also has to throw the points into the baskets. The straightforward

way of doing it is to check the point-basket pairs. This involves $2nT$ checks, which is worse than the naive n^2 algorithm! In order to overcome this, we take the k'th tensor power of the previous algorithm. That means throwing n^k points in $\{0,1\}^{kd}$ into T^k buckets, by dividing the coordinates into k blocks of size d. The success probability is S^k, the expected number of comparisons is at most T^k, but throwing the points into the baskets takes only an expected number of $2n^kT$ vector operations (of length kd). Hence the total expected number of vector operations is at most

$$T^k + 2n^kT \tag{22}$$

At last taking

$$k = \lceil 1/(1-p) \rceil \tag{23}$$

lets us approach the promised exponent $1/p$.

III. The Probabilistic Model

Definition 3.1: The pairwise independent identically distributed data model is the following. Let the sets

$$X_0 \subset \{0,1,\ldots,b_0-1\}^d, \quad X_1 \subset \{0,1,\ldots,b_1-1\}^d \tag{24}$$

of cardinalities $\#X_0 = n_0, \ \ \#X_1 = n_1$ be randomly constructed using the probability matrix

$$P = \begin{pmatrix} p_{00} & p_{01} & \cdots & p_{0\ b_1-1} \\ p_{10} & p_{11} & \cdots & p_{0\ b_1-1} \\ \vdots & \vdots & \ddots & \vdots \\ p_{b_0-1\ 0} & p_{b_0-1\ 1} & \cdots & p_{b_0-1\ b_1-1} \end{pmatrix} \tag{25}$$

$$p_{jk} \geq 0, \quad \sum_{j=0}^{b_0-1} \sum_{k=0}^{b_1-1} p_{jk} = 1 \tag{26}$$

The X_0 points are identically distributed pairwise independent Bernoulli random vectors, with

$$p_{j*} = \sum_{k=0}^{b_1-1} p_{jk} \tag{27}$$

probability that coordinate i has value j. The probability of a single point $x \in X_0$ is

$$p_{x*} = \prod_{i=1}^{d} p_{x_i *} \tag{28}$$

and the probability of a set $B_0 \subset X_0$ is of course

$$p_{B_0 *} = \sum_{x \in B_0} p_{x*} \tag{29}$$

Similarly X_1 is governed by $p_{*k} = \sum_{j=0}^{b_0-1} p_{jk}$ There is a special pair of X_0, X_1 points, uniformly chosen out of the $n_0 n_1$ possibilities. For that pair the probability that their i'th coordinates are j, k is p_{jk} and for $x_0 \in X_0,\ x_1 \in X_1$

$$p_{x_0 x_1} = \prod_{i=1}^{d} p_{x_{0,i} x_{1,i}} \tag{30}$$

Coding and information theory were initially developed for a similar model (with a probability vector instead of a probability matrix). Extension to non-uniform matrices, a stationary model with coordinate dependency, or continuous data is possible, as was done for coding and information theory.

IV. Comparison with the Indyk-Motwani Analysis

The Indyk-Motwani paper [6] introduces a metric based, worst case analysis. In general no average work upper bound can replace a worst case work upper bound, and the reverse holds for lower bounds. Still some comparison is unavoidable. Let us consider the uniform marginally Bernoulli(1/2) problem with $d \to \infty$. We saw that the classical approach requires $W \approx n^{\log_2 2/p}$, and have reduced it to $W \approx n^{\epsilon + 1/p}$. What is the Indyk-Motwani bound? The Hamming distance between two random points is approximately $d/2$ (the ratio to d tends to $1/2$ as d grows, according to the law of large numbers). The Hamming distance between two related points is approximately $(1-p)d$. Hence the distance ratio is $c = 1/(2-2p)$ and the Indyk-Motwani work is

$$W \approx n^{1+1/c} = n^{3-2p} \tag{31}$$

It can be argued that the drop in performance is offset by the lack of pairwise independence assumptions. The $n^{\frac{2}{1+e^{-1/c}}} = n^{\frac{2}{1+e^{2p-2}}}$ lower bound of Motwani, Naor and Panigrahy [8] is interesting, but increasing it to $n^{1/p}$ seems a challenge.

Now let us consider a typical sparse bits matrix: for a small ϵ let

$$P = \begin{pmatrix} 1-3\epsilon & \epsilon \\ \epsilon & \epsilon \end{pmatrix} \tag{32}$$

The standard bucketing approach is to arrange the coordinates randomly and hash each point by its first k 1'ns, where $k \approx -\ln n / \ln 2\epsilon$. The probability that two unrelated points fall into the same bucket is less than $(2\epsilon)^k \approx 1/n$, so the expected work per try is approximately n. The probability that the two related points fall into the same basket is at least

$$\begin{pmatrix} m \\ k \end{pmatrix} (1-3\epsilon)^{m-k} \epsilon^k = \begin{pmatrix} m \\ k \end{pmatrix} (1-3\epsilon)^{m-k} (3\epsilon)^k \cdot 3^{-k} \tag{33}$$

for any $m \geq k$ (consider the first m coordinates). Taking $m \approx k/3\epsilon$ shows that the success probability per try is at least approximately $3^{-k} \approx n^{\ln 3/\ln 2\epsilon}$. Hence in order to succeed we will make $n^{-\ln 3/\ln 2\epsilon}$ tries, and the total expected work is

$$W \approx n^{1+\frac{\ln 3}{\ln 1/2\epsilon}} \tag{34}$$

In contrast the Hamming distance between random points is approximately $2(1-2\epsilon)2\epsilon d$ and the Hamming distance between two related points is approximately $2\epsilon d$, so the Indyk-Motwani distance ratio is $c = 2(1-2\epsilon) \approx 2$ and

$$W \approx n^{1+1/c} \approx n^{3/2} \tag{35}$$

This worst case bound does not preclude the possibility that the random projections approach recommended for sparse data by Datar Indyk Immorlica and Mirrokni [4] performs better. Their optimal choice $r \to \infty$ results in a binary hash function $h(x) = \operatorname{sign}\left(\sum_{i=1}^{d} x_i C_i\right)$ where $(x_1, x_2, \ldots, x_d) \in X$ is a any point and $C_1, C_2, \ldots, C_d$ are independent Cauchy random variables (density $\frac{1}{\pi(1+z^2)}$). Both ± 1 values have probability $1/2$, so one has to concatenate $k \approx \log_2 n$

binary hash functions in order to determine a bucket. Now consider two related points. They will have approximately ϵd 1'ns in common, and each will have approximately ϵd 1'ns where the other has zeroes. The sum of ϵd independent Cauchy random variables has the same distribution as ϵd times a single Cauchy random variable, so the probability that the two related points get the same hash bit is approximately

$$\mathrm{Prob}\left\{\mathrm{sign}\left(C_1 + C_2\right) = \mathrm{sign}\left(C_1 + C_3\right)\right\} = 2/3 \tag{36}$$

Hence amount of work is large:

$$W \approx n(3/2)^k \approx n^{\log_2 3} \tag{37}$$

We have demonstrated that the probabilistic model adds to the current understanding of the approximate nearest neighbor problem. This is no surprise, since it is the standard model of information theory.

V. Bucketing Codes

Assume that there is enough information to identify the special pair. How much work is necessary? Comparing all $n_0 n_1$ point pairs suffice. All the effective known nearest neighbor algorithms are bucketing algorithms, so will limit ourselves to these. But what are bucketing algorithms? One could compute m_0, m_1 in some complicated way from the data, and then throw the m_0'th point of X_0 and the m_1'th point of X_1 into a single bucket. It is unlikely to work, but can you prove it? In order to disallow such knavery we will insist on data independent buckets. Most practical bucketing algorithms are data dependent. That is necessary because the data is used to construct (usually implicitly) a data model. We suspect that when the data model is known, there is little to be gained by making the buckets data dependent.

Definition 5.1: Assume the i.i.d. data model. A bucketing code is a set of T subset pairs

$$(B_{0,0}, B_{1,0}), \ldots, (B_{0,T-1}, B_{1,T-1}) \subset X_0 \times X_1$$

Its success probability is

$$S = p_{\cup_{t=0}^{T-1} B_{0,t} \times B_{1,t}} \tag{38}$$

and for any real numbers $n_0, n_1 > 0$ its work is

$$W = \sum_{t=0}^{T-1} \max\left(n_0 p_{B_{0,t}*}, n_1 p_{*B_{1,t}}, n_0 p_{B_{0,t}*} n_1 p_{*B_{0,t}}\right)$$

The meaning of success is obvious, but work has to be explained. In the above definition we consider n_0, n_1 to be the expected number of X_0, X_1 points, so they are not necessarily integers. The simplest implementation of a bucketing code is to store it as two point indexed arrays of lists. The first array of size b_0^d keeps for each point $x \in \{0, 1, \ldots, b_0 - 1\}^d$ the list of buckets (from 0 to $T-1$) which contain it. The second array of size b_1^d does the same for the $B_{1,t}$'s. When we are given X_0 and X_1 we look each element up, and accumulate pointers to it in a buckets array of k lists of pointers. Then we compare the pairs in each of the k buckets. Let us count the expected number of operations. The expected number of buckets containing any specific X_0 point is $\sum_{t=0}^{T-1} p_{B_{0,t}*}$, so the X_0 lookup involves an order of $n_0 + n_0 \sum_{t=0}^{T-1} p_{B_{0,t}*}$ operations. Similarly the X_1 lookup takes $n_1 + n_1 \sum_{t=0}^{T-1} p_{*B_{1,t}}$ The probability that a specific random pair falls into bucket t is $p_{B_{0,t}*} p_{*B_{1,t}}$, so the expected number of comparisons is $n_0 p_{B_{0,t}*} n_1 p_{*B_{1,t}}$ It all adds up to

$$n_0 + n_1 + \sum_{t=0}^{T-1} \left[n_0 p_{B_{0,t}*} + n_1 p_{*B_{1,t}} + n_0 p_{B_{0,t}*} n_1 p_{*B_{1,t}}\right] \le n_0 + n_1 + 3W \tag{39}$$

The fly in the ointment is that for even moderate dimension d the memory requirements of the previous algorithm are out of the universe. Hence it can be used only for small d. Higher dimensions can be handled by splitting them up into short blocks, or by more sophisticated coding algorithms.

VI. Basic Results

Definition 6.1: For any nonnegative matrix or vector R, and a probability matrix or vector P of the same dimensions $b_0 \times b_1$, let the extended Kullback-Leibler divergence be

$$K(R\|P) = \sum_{j=0}^{b_0-1} \sum_{k=0}^{b_1-1} r_{jk} \ln \frac{r_{jk}}{r_{**} p_{jk}} \ge 0 \tag{40}$$

where $r_{**} = \sum_{j=0}^{b_0-1} \sum_{k=0}^{b_1-1} r_{jk}$

Non-negativity follows from the well known inequality:

Lemma 6.1: For any nonnegative $q_0, q_1, \ldots, q_{b-1} \geq 0$, $p_0, p_1, \ldots, p_{b-1} \geq 0$

$$\sum_{j=0}^{b-1} q_j \ln \frac{q_j}{p_j} \geq q_* \ln \frac{q_*}{p_*} \tag{41}$$

where $q_* = \sum_{j=0}^{b-1} q_j$, $p_* = \sum_{j=0}^{b-1} p_j$

Definition 6.2: Suppose P is a probability matrix. We write that $\lambda_0, \lambda_1 \leq 1 \leq \lambda_0 + \lambda_1$ are P $\mathbf{sub-conjugate}$ to each other, denoted by $I(P, \lambda_0, \lambda_1, 1) = 0$, iff for any probability matrix Q of the same dimensions as P

$$K(Q_{\cdot\cdot} \| P_{\cdot\cdot}) \geq \lambda_0 K(Q_{\cdot *} \| P_{\cdot *}) + \lambda_1 K(Q_{*\cdot} \| P_{*\cdot}) \tag{42}$$

Explicitly

$$\sum_{j=0}^{b_0-1} \sum_{k=0}^{b_1-1} q_{jk} \ln \frac{q_{jk}}{p_{jk}} \geq \lambda_0 \sum_{j=0}^{b_0-1} q_{j*} \ln \frac{q_{j*}}{p_{j*}} + \lambda_1 \sum_{k=0}^{b_1-1} q_{*k} \ln \frac{q_{*k}}{p_{*k}} \tag{43}$$

where $q_{j*} = \sum_{k=0}^{b_1-1} q_{jk}$ etc. The set of P sub-conjugate pairs is convex by definition.

We will prove in the section VIII

Theorem 6.2: For any bucketing code with probability matrix P, set sizes n_0, n_1, success probability S and work W

$$W \geq S \sup_{\lambda_0, \lambda_1 \leq 1 \leq \lambda_0 + \lambda_1,\ I(P, \lambda_0, \lambda_1, 1) = 0} n_0^{\lambda_0} n_1^{\lambda_1} \tag{44}$$

The following inverse result is a special case of theorem 10.2

Theorem 6.3: For any probability matrices P, Q, a scalar $\epsilon > 0$ and large N there exists a bucketing code for matrix P, set sizes $n_0 = \lfloor N^{K(Q_{\cdot *} \| P_{\cdot *})} \rfloor$, $n_1 = \lfloor N^{K(Q_{*\cdot} \| P_{*\cdot})} \rfloor$, with success probability $S \geq 1 - \epsilon$ and work $W \leq N^{\epsilon + K(Q \| P)}$.

VII. An Example

Consider the classical matrix $P = \begin{pmatrix} p/2 & (1-p)/2 \\ (1-p)/2 & p/2 \end{pmatrix}$. Inserting $Q = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ into theorem 6.3 generates the well known $n_0 = n_1 \approx N^{\ln 2}$,$S \geq 1 - \epsilon$ and $W \leq N^{\epsilon + \ln 2/p}$.

The $Q \approx P$ neighborhood is important. Setting $q_{jk} = p_{jk} + \delta_{jk}$, $\delta_{jk} \to 0$, $\delta_{**} = 0$ results in $n_0 \approx N^{\sum_j \frac{\delta_{j*}^2}{2p_{j*}}}$, $n_1 \approx N^{\sum_k \frac{\delta_{*k}^2}{2p_{*k}}}$, $S \geq 1 - \epsilon$ and $W \leq N^{\epsilon + \sum_{jk} \frac{\delta_{jk}^2}{2p_{jk}}}$. Linear algebra shows that it is best to take $\delta_{00} = -\delta_{11} = \delta$, $\delta_{10} = -\delta_{01} = \alpha\delta$. Replacing N with N^{2/δ^2} and ϵ with $\epsilon\delta^2/2$

results in $n_0 \approx N^{(1-\alpha)^2}$, $n_1 \approx N^{(1+\alpha)^2}$, $S \geq 1-\epsilon$, $W \leq N^{\epsilon+1/p+\alpha^2/(1-p)}$. In particular for $\alpha = 0$ $n_0 = n_1 = n$, $S \geq 1-\epsilon$, $W \leq n^{\epsilon+1/p}$.

Is the exponent $1/p$ best possible? Theorem 6.2 reduces the optimality of $1/p$ to a single inequality:

Conjecture 7.1: For any $1/2 \leq p \leq 1$, $q_{00}, q_{01}, q_{10}, q_{11} \geq 0, q_{00}+q_{01}+q_{10}+q_{11} = 1$

$$2p\Big[q_{00}\ln\frac{2q_{00}}{p} + q_{01}\ln\frac{2q_{01}}{1-p} + q_{10}\ln\frac{2q_{10}}{1-p} + q_{11}\ln\frac{2q_{11}}{p}\Big] \geq \tag{45}$$

$$\geq (q_{00}+q_{01})\ln 2(q_{00}+q_{01}) + (q_{10}+q_{11})\ln 2(q_{10}+q_{11}) + \tag{46}$$

$$+(q_{00}+q_{10})\ln 2(q_{00}+q_{10}) + (q_{10}+q_{11})\ln 2(q_{10}+q_{11}) \tag{47}$$

Computer experimentation and critical point analysis leave no doubt that this inequality is valid. It is four dimensional, and keeping the marginal probabilities fixed shows that we can further restrict

$$(1-p)^2 q_{00}q_{11} = p^2 q_{01}q_{10} \tag{48}$$

A brute force proof is possible. Hopefully someone will find a clever proof.

Expressing N, α in terms of n_0, n_1 shows that we can do with $e^{\frac{\ln n_0 + \ln n_1 - 2(2p-1)\sqrt{\ln n_0 \ln n_1}}{4p(1-p)(1-\epsilon)}}$ comparisons. In particular when $n_0 = n_1^{(2p-1)^2-\epsilon}$, that asymmetric approximate nearest neighbor problem is solvable in linear time!

VIII. A Proof From The Book

In this section we will prove theorem 6.2.

Theorem 8.1: For any probability matrices P_1, P_2 and $\lambda_0, \lambda_1 \leq 1 \leq \lambda_0 + \lambda_1$

$$I(P_1,\lambda_0,\lambda_1,1) = I(P_2,\lambda_0,\lambda_1,1) = 0 \iff I(P_1\times P_2,\lambda_0,\lambda_1,1) = 0 \tag{49}$$

where $\times$ is tensor product.

Proof: Direction $\Leftarrow$ is obvious, so assume the left hand side. Denote $P = P_1 \times P_2$:

$$p_{j_1k_1j_2k_2} = p_{1,j_1k_1}p_{2,j_2k_2} \tag{50}$$

For any probability matrix $\{q_{j_1k_1j_2k_2}\}_{j_1k_1j_2k_2}$

$$\sum_{j_1k_1j_2k_2} q_{j_1k_1j_2k_2} \ln \frac{q_{j_1k_1j_2k_2}}{p_{1,j_1k_1}p_{2,j_2k_2}} = \sum_{j_1k_1} q_{j_1k_1**} \ln \frac{q_{j_1k_1**}}{p_{1,j_1k_1}} + \sum_{j_1k_1j_2k_2} q_{j_1k_1j_2k_2} \ln \frac{q_{j_1k_1j_2k_2}}{q_{j_1k_1**}p_{2,j_2k_2}} \tag{51}$$

Because $I(P_1, \lambda_0, \lambda_1, 1) = 0$

$$\sum_{j_1k_1} q_{j_1k_1**} \ln \frac{q_{j_1k_1**}}{p_{1,j_1k_1}} \geq \lambda_0 \sum_{j_1} q_{j_1***} \ln \frac{q_{j_1***}}{p_{1,j_1*}} + \lambda_1 \sum_{k_1} q_{*k_1**} \ln \frac{q_{*k_1**}}{p_{1,*k_1}} \tag{52}$$

Because $I(P_2, \lambda_0, \lambda_1, 1) = 0$

$$\sum_{j_2k_2} q_{j_1k_1j_2k_2}/q_{j_1k_1**} \ln \frac{q_{j_1k_1j_2k_2}/q_{j_1k_1**}}{p_{2,j_2k_2}} \geq \tag{53}$$

$$\geq \lambda_0 \sum_{j_2} q_{j_1k_1j_2*}/q_{j_1k_1**} \ln \frac{q_{j_1k_1j_2*}/q_{j_1k_1**}}{p_{2,j_2*}} + \lambda_1 \sum_{k_2} q_{j_1k_1*k_2}/q_{j_1k_1**} \ln \frac{q_{j_1k_1*k_2}/q_{j_1k_1**}}{p_{2,*k_2}} \tag{54}$$

$$\sum_{j_1k_1j_2k_2} q_{j_1k_1j_2k_2} \ln \frac{q_{j_1k_1j_2k_2}}{q_{j_1k_1**}p_{2,j_2k_2}} \geq \lambda_0 \sum_{j_1k_1j_2} q_{j_1k_1j_2*} \ln \frac{q_{j_1k_1j_2*}}{q_{j_1k_1**}p_{2,j_2*}} + \lambda_1 \sum_{j_1k_1k_2} q_{j_1k_1*k_2} \ln \frac{q_{j_1k_1*k_2}}{q_{j_1k_1**}p_{2,*k_2}} \tag{55}$$

so with help from lemma 6.1

$$\sum_{j_1k_1j_2k_2} q_{j_1k_1j_2k_2} \ln \frac{q_{j_1k_1j_2k_2}}{q_{j_1k_1**}p_{2,j_2k_2}} \geq \lambda_0 \sum_{j_1j_2} q_{j_1*j_2*} \ln \frac{q_{j_1*j_2*}}{q_{j_1***}p_{2,j_2*}} + \lambda_1 \sum_{k_1k_2} q_{*k_1*k_2} \ln \frac{q_{*k_1*k_2}}{q_{*k_1**}p_{2,*k_2}} \tag{56}$$

Together

$$\sum_{j_1k_1j_2k_2} q_{j_1k_1j_2k_2} \ln \frac{q_{j_1k_1j_2k_2}}{p_{1,j_1k_1}p_{2,j_2k_2}} \geq \lambda_0 \sum_{j_1j_2} q_{j_1*j_2*} \ln \frac{q_{j_1*j_2*}}{p_{1,j_1*}p_{2,j_2*}} + \lambda_1 \sum_{k_1k_2} q_{*k_1*k_2} \ln \frac{q_{*k_1*k_2}}{p_{1,*k_1}p_{2,*k_2}} \tag{57}$$

hence $I(P_1 \times P_2, \lambda_0, \lambda_1, 1) = 0$. ■

Theorem 8.2: For any $B_0 \subset \{0, 1, \ldots, b_0 - 1\}^d$, $B_1 \subset \{0, 1, \ldots, b_1 - 1\}^d$

$$p_{B_0B_1} \leq \min_{\lambda_0,\lambda_1 \leq 1 \leq \lambda_0+\lambda_1,\ I(P,\lambda_0,\lambda_1,0)=0} p_{B_0*}^{\lambda_0} p_{*B_1}^{\lambda_1} \tag{58}$$

Proof: Without restricting generality let $d = 1$. Inserting

$$q_{jk} = \begin{cases} \frac{p_{jk}}{p_{B_0B_1}} & j \in B_0, k \in B_1 \\ 0 & \text{otherwise} \end{cases} \tag{59}$$

into (42) proves the assertion. ■

Proof of **theorem 6.2**. *Proof:* Recall that the work is $W = \sum_i W_i$ where

$$W_i = \max\left(n_0 p_{B_{0,i}*},\ n_1 p_{*B_{1,i}},\ n_0 p_{B_{0,i}*} n_1 p_{*B_{0,i}}\right) \tag{60}$$

Our parameters satisfy

$$(\lambda_0, \lambda_1) \in \mathrm{Conv}(\{(1,0),(0,1),(1,1)\}) \tag{61}$$

hence

$$\ln W_i \geq \lambda_0 \ln(n_0 p_{B_{0,i}*}) + \lambda_1 \ln(n_1 p_{*B_{1,i}}) \tag{62}$$

$$W_i \geq n_0^{\lambda_0} n_1^{\lambda_1} p_{B_{0,i} B_{1,i}} \tag{63}$$

Now sum up. ■

IX. Bucketing Information

All the results of this section will be proven in appendix I.

Definition 9.1: Suppose P is a probability matrix. The **bucketing information** function is for $\mu \geq 0$

$$I(P, \lambda_0, \lambda_1, \mu) = \max_{\substack{\{r_{i,jk} \geq 0\} \\ 0 \leq i < b_0 b_1 \\ 0 \leq j < b_0 \\ 0 \leq k < b_1 \\ r_{*,**} = 1}} \left[\lambda_0 \sum_{i=0}^{b_0 b_1 - 1} K(R_{i,\cdot *} \| P_{\cdot *}) + \lambda_1 \sum_{i=0}^{b_0 b_1 - 1} K(R_{i,* \cdot} \| P_{* \cdot}) + \right.$$
$$\left. +(1-\mu) K(R_{*,\cdot\cdot} \| P_{\cdot\cdot}) - \sum_{i=0}^{b_0 b_1 - 1} K(R_{i,\cdot\cdot} \| P_{\cdot\cdot})\right]$$

Explicitly$r_{i,j*} = \sum_{k=0}^{b_1-1} r_{i,jk}$, $K(R_{i,\cdot *} \| P_{\cdot *}) = \sum_{j=0}^{b_0-1} r_{i,j*} \ln \frac{r_{i,j*}}{r_{i,**} p_{j*}}$ etc.

Lemma 9.1: For any probability matrix P and $0 \leq \mu \leq 1$ the sums in definition 9.1 can be restricted to a single term, i.e.

$$I(P, \lambda_0, \lambda_1, \mu) = \max_Q \left[\lambda_0 K(Q_{\cdot *} \| P_{\cdot *}) + \lambda_1 K(Q_{* \cdot} \| P_{* \cdot}) - \mu K(Q_{\cdot\cdot} \| P_{\cdot\cdot})\right] \tag{64}$$

where Q ranges over all probability matrices. For any $\mu \geq 0$, not restricting the number of terms

i in definition 9.1 does not change I. It can be rewritten as

$$I(P,\lambda_0,\lambda_1,\mu)=\max_Q\left[(1-\mu)K(Q_{\cdot\cdot}\|P_{\cdot\cdot})+\max_{(Q,y)\in\mathrm{Conv}(G(P,\lambda_0,\lambda_1))}y\right] \tag{65}$$

where Conv is the convex hull and

$$G(P,\lambda_0,\lambda_1)=\{(Q,\ \lambda_0K(Q_{\cdot *}\|P_{\cdot *})+\lambda_1K(Q_{*\cdot}\|P_{*\cdot})-K(Q_{\cdot\cdot}\|P_{\cdot\cdot}))\}_Q \tag{66}$$

From now on when dealing with the bucketing information function, we will denote $\sum_i$ without worrying about the number of indices.

Lemma 9.2: For any probability matrix P and $\mu\geq 0$ the bucketing information function $I(P,\lambda_0,\lambda_1,\mu)$ is nonnegative, convex, monotonically nondecreasing in λ_0,λ_1 and monotonically non-increasing in μ. Special values are

$$I(P,\lambda_0,\lambda_1,\mu)=\mu I(P,\lambda_0/\mu,\lambda_1/\mu,1)\quad 0<\mu\leq 1 \tag{67}$$

$$I(P,\lambda_0,\lambda_1,\mu)=0\iff\forall Q,\ \min(\mu,1)K(Q_{\cdot\cdot}\|P_{\cdot\cdot}))\geq\lambda_0K(Q_{\cdot *}\|P_{\cdot *})+\lambda_1K(Q_{*\cdot}\|P_{*\cdot}) \tag{68}$$

$$I(P,\lambda_0,\lambda_1,\mu)=0\qquad 0\leq\lambda_0,\lambda_1\quad\lambda_0+\lambda_1\leq\min(\mu,1) \tag{69}$$

$$I(P,1,1,\mu)=\max_{0\leq j<b_0,0\leq k<b_1}\ln\frac{(p_{jk})^\mu}{p_{j*}p_{*k}}\qquad 0\leq\mu\leq 1 \tag{70}$$

$$I(P,1,1,\mu)=(\mu-1)\ln\sum_{j=0}^{b_0-1}\sum_{k=0}^{b_1-1}p_{jk}\left(\frac{p_{jk}}{p_{j*}p_{*k}}\right)^{\frac{1}{\mu-1}}\qquad\mu\geq 1 \tag{71}$$

$$I(P,1,1,\infty)=I(P)=\sum_{j=0}^{b_0-1}\sum_{k=0}^{b_1-1}p_{jk}\ln\frac{p_{jk}}{p_{j*}p_{*k}} \tag{72}$$

Theorem 9.3: For any probability matrices P_1,P_2 and $\mu\geq 0$

$$I(P_1\times P_2,\lambda_0,\lambda_1,\mu)=I(P_1,\lambda_0,\lambda_1,\mu)+I(P_2,\lambda_0,\lambda_1,\mu) \tag{73}$$

X. Bucketing Codes and Information

All the results of this section will be proven in appendix II.

Theorem 10.1: For any bucketing code with probability matrix $P_1\times P_2\times\cdots\times P_{\tilde{d}}$, dimension

$d=1$, set sizes n_0, n_1, success probability S and work W

$$\ln W \geq \sup_{\lambda_0,\lambda_1 \leq 1 \leq \lambda_0+\lambda_1,\ \mu \geq 0} \left[\lambda_0 \ln n_0 + \lambda_1 \ln n_1 + \mu \ln S - \sum_{i=1}^{\tilde{d}} I(P_i, \lambda_0, \lambda_1, \mu) \right] \tag{74}$$

Definition 10.1: Assume the i.i.d. data model with probability matrix P. Suppose there exists a d dimensional bucketing code such that for the expected numbers n_0, n_1 of X_0, X_1 points it has success probability S and work W. Then for any real numbers $0 \leq \tilde{S} \leq S,\ \tilde{W} \geq W$ we say that $(P, d, n_0, n_1, \tilde{S}, \tilde{W})$ is $\mathbf{attainable}$. Define the set of $\mathbf{log-attainable}$ parameters to be

$$D(P) = \left\{ \frac{1}{d}(\ln n_0, \ln n_1, -\ln S, \ln W) \ \middle|\ (P, d, n_0, n_1, S, W) \text{ is attainable} \right\} \tag{75}$$

Normalizing by d is awkward in the infinite data case $d = \infty$. There it makes sense to consider the $\mathbf{log-attainable\ cone}$

$$D_0(P) = \mathrm{Cone}(D(P)) = \cup_{\alpha \geq 0} \alpha D(P) \tag{76}$$

Theorem 10.1 is asymptotically tight in the following sense:

Theorem 10.2: For any probability matrix P the closure of its log-attainable set is

$$D^c(P) \ = \ \{(m_0, m_1, s, w) \mid s \geq 0, \tag{77}$$

$$\forall\ \lambda_0, \lambda_1 \leq 1 \leq \lambda_0 + \lambda_1,\ \mu \geq 0 \quad w \geq \lambda_0 m_0 + \lambda_1 m_1 - \mu s - I(P, \lambda_0, \lambda_1, \mu)\} \tag{78}$$

Equivalently

$$D^c(P) \ = \ D(0) + \mathrm{Conv}\Big(\Big\{\Big(\sum_i K(R_{i,\cdot *} \| P_{\cdot *}),\ \sum_i K(R_{i,*\cdot} \| P_{*\cdot}),\ K(R_{*,\cdot\cdot} \| P_{\cdot\cdot}), \tag{79}$$

$$-K(R_{*,\cdot\cdot} \| P_{\cdot\cdot}) + \sum_i K(R_{i,\cdot\cdot} \| P_{\cdot\cdot})\Big)\Big\}_{r_{i,jk} \geq 0,\ r_{*,**}=1}\Big) \tag{80}$$

where $D(0)$ is the common core

$$D(0) = \mathrm{ConvCone}(\{(1,0,0,1), (0,1,0,1), (0,0,1,0), (0,0,0,1), (-1,-1,0,-1)\}) \tag{81}$$

For the unlimited data case $d \to \infty$

$$D_0^c(P) = \{(m_0, m_1, s, w) \mid s \geq 0\} \cap \tag{82}$$

$$\cap[D_0(0) + \mathrm{ConvCone}(\{(K(Q_{\cdot *}\|P_{\cdot *}),\ K(Q_{*\cdot}\|P_{*\cdot}),\quad K(Q_{\cdot\cdot}\|P_{\cdot\cdot}),\ 0)\}_Q)] \tag{83}$$

where $D_0(0)$ is the extended common core

$$D_0(0) = D(0) + \mathrm{Cone}(\{(0, 0, -1, 1)\}) \tag{84}$$

and Q runs over all $b_0 \times b_1$ probability matrices.

In light of theorem 10.2, theorem 9.3 can be recast as

Theorem 10.3: For any probability matrices P_1, P_2 $\;D^c(P_1 \times P_2) = D^c(P_1) + D^c(P_2)$

XI. Conclusion

We consider the approximate nearest neighbor problem in a probabilistic setting. Using several coordinates at once enables asymptotically better approximate nearest neighbor algorithms than using them one at a time. The performance is bounded by, and tends to, a newly defined bucketing information function. Thus bucketing coding and information theory play the same role for the approximate nearest neighbor problem that Shannon's coding and information theory play for communication.

References

[1] N. Alon Private Communication.

[2] A. Andoni, P. Indyk *Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions* FOCS 2006.

[3] A. Broder. *Identifying and Filtering Near-Duplicate Documents* Proc. FUN, 1998.

[4] M. Datar, P. Indyk, N. Immorlica and V. Mirrokni *Locality-Sensitive Hashing Scheme Based on p-Stable Distributions* Proc. Sympos. on Computational Geometry, 2004.

[5] C.Gennaro, P.Savino and P.Zezula *Similarity Search in Metric Databases through Hashing* Proc. ACM workshop on multimedia, 2001.

[6] P. Indyk and R. Motwani. *Approximate Nearest Neighbor: Towards Removing the Curse of Dimensionality* Proc. 30th Annu. ACM Sympos. Theory Comput., 1998.

[7] R.M. Karp, O. Waarts, and G. Zweig. *The Bit Vector Intersection Problem* Proc. 36th Annu. IEEE Sympos. Foundations of Computer Science, pp. 621-630, 1995.

[8] R. Motwani, A. Naor and R. Panigrahy *Lower Bounds on Locality Sensitive Hashing* SCG'06

[9] *A Heterogeneous High Dimensional Approximate Nearest Neighbor Algorithm* To be Published.

Appendix I

Bucketing Information Proofs

Proof of **Lemma 9.1**. *Proof:* Lemma (6.1) implies that

$$K(R_{*,\cdot\cdot}\|P_{\cdot\cdot}) \leq \sum_{i=0}^{b_0 b_1 - 1} K(R_{i,\cdot\cdot}\|P_{\cdot\cdot}) \tag{85}$$

so for $0 \leq \mu \leq 1$

$$(1-\mu)K(R_{*,\cdot\cdot}\|P_{\cdot\cdot}) - \sum_{i=0}^{b_0 b_1 - 1} K(R_{i,\cdot\cdot}\|P_{\cdot\cdot}) \leq -\mu \sum_{i=0}^{b_0 b_1 - 1} K(R_{i,\cdot\cdot}\|P_{\cdot\cdot}) \tag{86}$$

and only one i is necessary. The connection between definition 9.1 and (65) is through $r_i = r_{i,**}$, $q_{i,jk} = \frac{r_{i,jk}}{r_{i,**}}$

$$I(P,\lambda_0,\lambda_1,\mu) = \max_{\substack{\{r_i, Q_i\}_i \\ r_* = 1}} \left[\sum_{i=0}^{b_0 b_1 - 1} r_i \Big[\lambda_0 K(Q_{i,\cdot *}\|P_{\cdot *}) + \lambda_1 K(Q_{i,*\cdot}\|P_{*\cdot}) + \right. \tag{87}$$

$$\left. + (1-\mu) K\Big(\sum_i r_i Q_{i,\cdot\cdot} \Big\| P_{\cdot\cdot} \Big) - K(Q_{i,\cdot\cdot}\|P_{\cdot\cdot}) \Big] \right] \tag{88}$$

The set G is $b_0 b_1$ dimensional, so by Caratheodory's theorem any point on the boundary of its convex hull is a convex combination of $b_0 b_1$ G points. ■

Proof of **lemma 9.2**. *Proof:* Non-negativity follows by taking $Q = P$. Monotonicity ,convexity and (67) are by definition.

When $0 \leq \mu \leq 1$ (64) is valid and (68) is clear. When $\mu \geq 1$

$$I(P,\lambda_0,\lambda_1,\mu) \leq \max_R \sum_{i=0}^{b_0 b_1 - 1} \Big[\lambda_0 K(R_{i,\cdot *}\|P_{\cdot *}) + \lambda_1 K(R_{i,*\cdot}\|P_{*\cdot}) - K(R_{i,\cdot\cdot}\|P_{\cdot\cdot}) \Big] \tag{89}$$

so direction $\Leftarrow$ of (68) is true. On the other hand assume that for some Q

$$K(Q_{\cdot\cdot}\|P_{\cdot\cdot}) < \lambda_0 K(Q_{\cdot *}\|P_{\cdot *}) + \lambda_1 K(Q_{*\cdot}\|P_{*\cdot}) \tag{90}$$

Inserting $r_{0,jk} = \epsilon q_{jk},\ r_{1,jk} = p_{jk} - \epsilon q_{jk}$ into definition 9.1 gives

$$I(P,\lambda_0,\lambda_1,\mu) \geq \epsilon \left[\lambda_0 K(Q_{\cdot *}\|P_{\cdot *}) + \lambda_1 K(Q_{*\cdot}\|P_{*\cdot}) - K(Q_{\cdot\cdot}\|P_{\cdot\cdot}) \right] + \tag{91}$$

$$+(1-\epsilon)\left[\lambda_0 K(\tilde{P}_{\cdot *}\|P_{\cdot *}) + \lambda_1 K(\tilde{P}_{*\cdot}\|P_{*\cdot}) - K(\tilde{P}_{\cdot\cdot}\|P_{\cdot\cdot})\right] \tag{92}$$

where $\tilde{P} = (P-\epsilon Q)/(1-\epsilon) = P + \epsilon(P-Q)/(1-\epsilon)$. The Kullback-Leibler divergence between $\tilde{P}$ and P is second order in ϵ, and the same holds for their marginal vectors. Hence for a small $\epsilon > 0$ $I(P,\lambda_0,\lambda_1,\mu) > 0$, and the proof of (68) is done.

Lemma 6.1 implies

$$K(R_{i,\cdot *}\|P_{\cdot *}), K(R_{i,*\cdot}\|P_{*\cdot}) \le K(R_{i,\cdot\cdot}\|P_{\cdot\cdot}) \tag{93}$$

so (69) follows from (68).

Now to $\lambda_0 = \lambda_1 = 1$. We want to maximize

$$\sum_i \left[K(R_{i,\cdot *}\|P_{\cdot *}) + K(R_{i,*\cdot}\|P_{*\cdot}) - K(R_{i,\cdot\cdot}\|P_{\cdot\cdot})\right] = \sum_{jk} r_{*,jk} \ln \frac{p_{jk}}{p_{j*}p_{*k}} - \sum_{ijk} r_{i,jk} \ln \frac{r_{i,**}r_{i,jk}}{r_{i,j*}r_{i,*k}}$$

The rightmost sum is nonnegative, and for any $\{r_{*,jk}\}_{jk}$ it can be made 0 by choosing

$$r_{i,jk} = \begin{cases} r_{*,jk} & i = j + b_0 k \\ 0 & \text{otherwise} \end{cases} \tag{94}$$

Hence we want to maximize

$$\sum_{jk} r_{*,jk} \ln \frac{(p_{jk})^\mu}{p_{j*}p_{*k}} + (1-\mu)\sum_{jk} r_{*,jk} \ln r_{*,jk} \tag{95}$$

When $0 \le \mu \le 1$ both sums can be simultaneously maximized by concentrating r in one place. When $\mu \ge 1$ the maximized function is concave in $\{r_{*,jk}\}_{jk}$, and Lagrange multipliers reveal the optimal choice

$$r_{*,jk} = \frac{\left(\frac{(p_{jk})^\mu}{p_{j*}p_{*k}}\right)^{\frac{1}{\mu-1}}}{\sum_{\tilde{j}\tilde{k}}\left(\frac{(p_{\tilde{j}\tilde{k}})^\mu}{p_{\tilde{j}*}p_{*\tilde{k}}}\right)^{\frac{1}{\mu-1}}} \tag{96}$$

■

Proof of **theorem 9.3**. *Proof:* Obviously $I(P_1 \times P_2, \lambda_0, \lambda_1, \mu) \ge I(P_1, \lambda_0, \lambda_1, \mu) + I(P_2, \lambda_0, \lambda_1, \mu)$. The other direction is the challenge. Denote $P = P_1 \times P_2$:

$$p_{j_1 k_1 j_2 k_2} = p_{1,j_1 k_1} p_{2,j_2 k_2} \tag{97}$$

For any $\{r_{i,j_1j_2k_1k_2}\}_{i,j_1j_2k_1k_2}$

$$(\mu-1)K(R_{*,\dots\dots}\|P_{\dots\dots}) + \sum_i K(R_{i,\dots\dots}\|P_{\dots\dots}) =$$

$$= (\mu-1)\sum_{j_1k_1} r_{*,j_1k_1**}\ln\frac{r_{*,j_1k_1**}}{p_{1,j_1k_1}} + \sum_{i,j_1k_1} r_{i,j_1k_1**}\ln\frac{r_{i,j_1k_1**}}{r_{i,****}p_{1,j_1k_1}} +$$
$$+(\mu-1)\sum_{j_1k_1j_2k_2} r_{*,j_1k_1j_2k_2}\ln\frac{r_{*,j_1k_1j_2k_2}}{r_{*,j_1k_1**}p_{2,j_2k_2}} + \sum_{i,j_1k_1j_2k_2} r_{i,j_1k_1j_2k_2}\ln\frac{r_{i,j_1k_1j_2k_2}}{r_{i,j_1k_1**}p_{2,j_2k_2}}$$

By definition

$$(\mu-1)\sum_{j_1k_1} r_{*,j_1k_1**}\ln\frac{r_{*,j_1k_1**}}{p_{1,j_1k_1}} + \sum_{i,j_1k_1} r_{i,j_1k_1**}\ln\frac{r_{i,j_1k_1**}}{r_{i,****}p_{1,j_1k_1}} \geq$$
$$\geq \lambda_0\sum_{i,j_1} r_{i,j_1***}\ln\frac{r_{i,j_1***}}{r_{i,****}p_{1,j_1*}} + \lambda_1\sum_{i,k_1} r_{i,*k_1**}\ln\frac{r_{i,*k_1**}}{r_{i,****}p_{1,*k_1}} - I(P_1,\lambda_0,\lambda_1,\mu)$$

$$(\mu-1)\sum_{j_2k_2} r_{*,j_1k_1j_2k_2}\ln\frac{r_{*,j_1k_1j_2k_2}}{r_{*,j_1k_1**}p_{2,j_2k_2}} + \sum_{i,j_2k_2} r_{i,j_1k_1j_2k_2}\ln\frac{r_{i,j_1k_1j_2k_2}}{r_{i,j_1,k_1**}p_{2,j_2k_2}} \geq$$
$$\geq \lambda_0\sum_{i,j_2} r_{i,j_1k_1j_2*}\ln\frac{r_{i,j_1k_1j_2*}}{r_{i,j_1k_1**}p_{2,j_2*}} + \lambda_1\sum_{i,k_2} r_{i,j_1k_1*k_2}\ln\frac{r_{i,j_1k_1*k_2}}{r_{i,j_1,k_1**}p_{2,*k_2}} -$$
$$-r_{*,j_1k_1**}I(P_2,\lambda_0,\lambda_1,\mu)$$

so with help from lemma 6.1

$$(\mu-1)\sum_{j_1k_1j_2k_2} r_{*,j_1k_1j_2k_2}\ln\frac{r_{*,j_1k_1j_2k_2}}{r_{*,j_1k_1**}p_{2,j_2k_2}} + \sum_{i,j_1k_1j_2k_2} r_{i,j_1k_1j_2k_2}\ln\frac{r_{i,j_1k_1j_2k_2}}{r_{i,j_1,k_1**}p_{2,j_2k_2}} \geq$$
$$\geq \lambda_0\sum_{i,j_1j_2} r_{i,j_1*j_2*}\ln\frac{r_{i,j_1*j_2*}}{r_{i,j_1***}p_{2,j_2*}} + \lambda_1\sum_{i,k_1k_2} r_{i,*k_1*k_2}\ln\frac{r_{i,*k_1*k_2}}{r_{i,*,k_1**}p_{2,*k_2}} - I(P_2,\lambda_0,\lambda_1,\mu)$$

Together

$$(\mu-1)K(R_{*,\dots\dots}\|P_{\dots\dots}) + \sum_i K(R_{i,\dots\dots}\|P_{\dots\dots}) \geq$$
$$\geq \lambda_0\sum_i K(R_{i,\cdot*\cdot*}\|P_{\cdot*\cdot*}) + \lambda_1\sum_i K(R_{i,*\cdot*\cdot}\|P_{*\cdot*\cdot}) - I(P_1,\lambda_0,\lambda_1,\mu) - I(P_2,\lambda_0,\lambda_1,\mu)$$

Notice that we have used the fact that for $0\leq\mu\leq 1$ there is only one i. ■

Appendix II

Bucketing Codes and Information Proofs

Proof of **theorem 10.1**. *Proof:* Without restricting generality let $\nu = 1$. Let $(B_{0,0}, B_{1,0}), \cdots, (B_{0,T-1}, B_{1,T-1})$ be subset pairs. Denote

$$B_i = B_{0,i} \times B_{1,i} \setminus \bigcup_{t=0}^{i-1} B_{0,t} \times B_{1,t} \tag{98}$$

so the success probability is $S = \sum_i p_{B_i}$ Insert

$$r_{i,jk} = \begin{cases} \frac{p_{jk}}{S} & (j,k) \in B_i \\ 0 & \text{otherwise} \end{cases} \tag{99}$$

into definition 9.1. Lemma 6.1 implies

$$K(R_{i,\cdot *} \| P_{\cdot *}) = \sum_{j \in B_{0,i}} r_{i,j*} \ln \frac{r_{i,j*}}{r_{i,**} p_{j*}} \geq -r_{i,**} \ln p_{B_{0,i}*} \tag{100}$$

Similarly

$$K(R_{i,*\cdot} \| P_{*\cdot}) \geq -r_{i,**} \ln p_{*B_{1,i}} \tag{101}$$

$$\sum_i \left[\lambda_0 K(R_{i,\cdot *} \| P_{\cdot *}) + \lambda_1 K(R_{i,*\cdot} \| P_{*\cdot})\right] \geq -\sum_i r_{i,**} \left(\lambda_0 \ln p_{B_{0,i}*} + \lambda_1 \ln p_{*B_{1,i}}\right) \tag{102}$$

Recall that the work is $W = \sum_i W_i$ where

$$W_i = \max\left(n_0 p_{B_{0,i}*},\ n_1 p_{*B_{1,i}},\ n_0 p_{B_{0,i}*} n_1 p_{*B_{0,i}}\right) \tag{103}$$

Our parameters satisfy

$$(\lambda_0, \lambda_1) \in \mathrm{Conv}(\{(1,0), (0,1), (1,1)\}) \tag{104}$$

hence

$$\ln W_i \geq \lambda_0 \ln(n_0 p_{B_{0,i}*}) + \lambda_1 \ln(n_1 p_{*B_{1,i}}) \tag{105}$$

$$-\lambda_0 \ln p_{B_{0,i}*} - \lambda_1 \ln p_{*B_{1,i}} \geq \lambda_0 \ln n_0 + \lambda_1 \ln n_1 - \ln W_i \tag{106}$$

Clearly

$$K(R_{*,\cdot\cdot} \| P_{\cdot\cdot}) = -\ln S \tag{107}$$

$$K(R_{i,\cdot\cdot}\|P_{\cdot\cdot}) = -\sum_{ijk} r_{i,jk}\ln(r_{i,**}S) = -\ln S - \sum_i r_{i,**}\ln r_{i,**} \tag{108}$$

Now all the pieces come together:

$$I(P,\lambda_0,\lambda_1,\mu) \geq \lambda_0\ln n_0 + \lambda_1\ln n_1 - \sum_i r_{i,**}\ln W_i + \mu\ln S + \sum_i r_{i,**}\ln r_{i,**} =$$
$$= \lambda_0\ln n_0 + \lambda_1\ln n_1 + \mu\ln S + \sum_i r_{i,**}\ln\frac{r_{i,**}}{W_i}$$

Another call of duty for lemma 6.1 produces

$$\sum_i r_{i,**}\ln\frac{r_{i,**}}{W_i} \geq -\ln W \tag{109}$$

■

Lemma 2.1: Suppose that

$$(P_1, d, n_{0,1}, n_{1,1}, S_1, W_1),\ (P_2, d, n_{0,2}, n_{1,2}, S_2, W_2) \tag{110}$$

are attainable. Then

$$(P_1\times P_2, d, n_{0,1}n_{0,2}, n_{1,1}n_{1,2}, S_1S_2, W_1W_2) \tag{111}$$

is attainable, where $\times$ is tensor product. In particular when $P_1 = P_2 = P$ for any $k_1, k_2 \geq 0$ we attain

$$(P, (k_1+k_2)d, n_{0,1}^{k_1}n_{0,2}^{k_2}, n_{1,1}^{k_1}n_{1,2}^{k_2}, S_1^{k_1}S_2^{k_2}, W_1^{k_1}W_2^{k_2}) \tag{112}$$

In particular the closure of the log-attainable set $D^c(P)$ is convex.

Proof: Tensor product the codes. ■

Lemma 2.2: Suppose that

$$(P, d_1, n_0, n_1, S_1, W_1),\ (P, d_2, n_0, n_1, S_2, W_2) \tag{113}$$

are attainable. Then

$$(P, d_1+d_2, n_0, n_1, S_1+S_2-S_1S_2, W_1+W_2) \tag{114}$$

is attainable. In particular for any $S_1 \leq \tilde{S}_1 \leq 1$

$$(\ln n_0, \ln n_1, -\ln S_1/\tilde{S}_1, \ln W_1/\tilde{S}_1) \in D_0^c(P) \tag{115}$$

Proof: Concatenating the codes shows the first claim. Concatenating T times the k'th tensor power of the first code shows that

$$\left(P, Td_1^k, n_0^k, n_1^k, 1 - \left(1 - S_1^k\right)^T, TW_1^k\right) \tag{116}$$

is attainable. Taking $T = \lceil \tilde{S}_1^{-k} \rceil$ and letting $k \to \infty$ finishes the proof. ■

Proof of **theorem 10.2**. *Proof:* First let us show that the two representations are equivalent. Denote the right hand side of (79) by E. It is the dual of its dual:

$$\begin{aligned} E = \{(m_0, m_1, s, w) \mid \alpha_0 m_0 + \alpha_1 m_1 - \beta s - \gamma w \leq 1 \\ \forall \alpha_0, \alpha_1, \beta, \gamma, R \ \text{ such that } \ \alpha_0, \alpha_1 \leq \gamma \leq \alpha_0 + \alpha_1, \ \beta, \gamma \geq 0, \\ \alpha_0 \sum_i K(R_{i,\cdot *} \| P_{\cdot *}) + \alpha_1 \sum_i K(R_{i,*\cdot} \| P_{*\cdot}) + (\gamma - \beta) K(R_{*,\cdot\cdot} \| P_{\cdot\cdot}) - \gamma \sum_i K(R_{i,\cdot\cdot} \| P_{\cdot\cdot}) \leq 1\} \end{aligned}$$

When $\gamma = 0$ it forces $\alpha_0 = \alpha_1 = 0$ and we are left with $-\beta s \leq 1$ for all $\beta \geq 0$, i.e. $s \geq 0$. When $\gamma > 0$ we can divide by it, denote $\lambda_0 = \alpha_0/\gamma, \lambda_1 = \alpha_1/\gamma, \mu = \beta/\gamma$ and find that $1/\gamma \geq I$ so E equals the right hand side of (77).

Theorem 10.1 implies that $D^c(P) \subset E$. We will prove the inverse inclusion by construction. The single big bags pair code

$$B_0 = \{0, 1, \ldots, b_0 - 1\}, \ B_1 = \{0, 1, \ldots, b_1 - 1\} \tag{117}$$

shows that $D(0) \subset D(P)$. Now let $\{r_{i,jk}\}_{ijk}$ attain the bucketing information value I. For dimension d choose integers $\{d_{i,jk}\}_{ijk}$ such that $d_{*,**} = d$ and

$$r_{i,jk} d - 1 < d_{i,jk} < r_{i,jk} d + 1 \tag{118}$$

Let us define a bucket pair

$$B_{0,0} = \left\{ x_0 \ \middle| \ \forall ij \ \sum_{l=c_i+1}^{c_{i+1}} (x_{0,l} == j) \ = \ d_{i,j*} \right\} \tag{119}$$

$$B_{0,1} = \left\{ x_1 \;\middle|\; \forall ik \sum_{l=c_i+1}^{c_{i+1}} (x_{1,l} == k) \;=\; d_{i,*k} \right\} \tag{120}$$

where $c_i = \sum_{l=0}^{i-1} d_{i,**}$ In words we want x_0 to contain exactly $d_{0,j*}$ j-values in its first $d_{0,**}$ coordinates, etc. The bucket size is

$$p_{B_{0,0*}} = \prod_i \left[\frac{d_{i,**}!}{\prod_j d_{i,j*}!} \prod_j p_{j*}^{d_{i,j*}} \right] \tag{121}$$

$$p_{*B_{0,1}} = \prod_i \left[\frac{d_{i,**}!}{\prod_k d_{i,*k}!} \prod_k p_{*k}^{d_{i,*k}} \right] \tag{122}$$

Let us add $T-1$ similar buckets. They are generated by randomly permuting the coordinates $1, 2, \ldots, d$. Let $n_0 = 1/p_{B_{0,0*}}$, $n_1 = 1/p_{*B_{0,1}}$ so that the work is $W = T$. A lower bound of the average success probability of this random bucketing code is

$$\mathrm{E}[S] \geq U \left[1 - (1 - V/U)^T \right] \tag{123}$$

where

$$U = \frac{d!}{\prod_{jk} d_{*,jk}!} \prod_{jk} p_{jk}^{d_{*,jk}} \tag{124}$$

is the probability that the special pair obtains coordinate pair (j,k) exactly $d_{*,jk}$ times, and

$$V = \prod_i \left[\frac{d_{i,**}!}{\prod_{jk} d_{i,jk}!} \prod_{jk} p_{jk}^{d_{i,jk}} \right] \tag{125}$$

is the probability that the special pair obtains coordinate pair (j,k) exactly $d_{i,jk}$ times in coordinate subset number i. Of course there exists a deterministic code at least as successful as the average code.

It is reasonable to take $T = \lceil U/V \rceil$. Stirling's approximation implies

$$\lim_{d \to \infty} \frac{1}{d} \ln n_0 = \sum_{ij} r_{i,j*} \ln \frac{r_{i,j*}}{r_{i,**} p_{j*}} \tag{126}$$

$$\lim_{d \to \infty} \frac{1}{d} \ln n_1 = \sum_{ik} r_{i,*k} \ln \frac{r_{i,*k}}{r_{i,**} p_{*k}} \tag{127}$$

$$\lim_{d\to\infty} \frac{-1}{d} \ln U = \sum_{jk} r_{*,jk} \ln \frac{r_{*,jk}}{p_{jk}} \tag{128}$$

$$\lim_{d\to\infty} \frac{-1}{d} \ln V = \sum_{ijk} r_{i,jk} \ln \frac{r_{i,jk}}{r_{i,**}p_{jk}} \tag{129}$$

Hence

$$\liminf_{d\to\infty} \frac{1}{d}(\lambda_0 \ln n_0 + \lambda_1 \ln n_1 + \mu \ln S - \ln W) \geq \tag{130}$$

$$\geq \lim_{d\to\infty} \frac{1}{d}(\lambda_0 \ln n_0 + \lambda_1 \ln n_1 + (\mu - 1) \ln U + \ln V) = I \tag{131}$$

There remains the unlimited data formula (82). Lemmas 2.2 shows that

$$D_0^c(P) = \tilde{D}_0(P) \cap \{(m_0, m_1, s, w \mid s \geq 0\} \tag{132}$$

$$\tilde{D}_0(P) = D_0^c(0) + \mathrm{Cone}(\{(0, 0, -1, 1)\}) \tag{133}$$

Clearly $\tilde{D}_0(P)$ is convex, contains the origin, and any point $(\alpha_0, \alpha_1, \beta, \gamma)$ in its dual satisfies $\beta \leq \gamma$. Hence $\mu = \beta/\gamma \leq 1$ so by lemma 9.1 only one i term is needed, as long as we use the full $D_0(0)$. ■